PENENTUAN KERELEVANAN HALAMAN WEB DENGAN MENGGUNAKAN ALGORITMA SALSA PADA INFORMATION RETRIEVAL

DECISION MAKING IN THE RELEVANCY WITHIN WEB PAGING BY USING SALSA ALGORITHM ON INFORMATION RETRIEVAL

Zarot Hendra Siagian[1], Yanuar Firdaus A.w.[2], Warih Maharani[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Proses pencarian dan pengurutan dokumen halaman web di dalam mesin pencari terbagi ke dalam 2 bagian yaitu dengan menggunakan text-based analysis dan link-based analysis. Proses text-based analysis ialah proses perangkingan dokumen halaman web berdasarkan teks yang ada di dokumen halaman web tersebut. Pada proses pencarian dengan menggunakan konsep ini memiliki beberapa kekurangan diantaranya untuk mencari dokumen yang relevan sesuai dengan keinginan user maka user diharapkan mampu memasukkan query setepat mungkin dengan term yang ada di dalam dokumen.

Pencarian dengan menggunakan konsep link-based analysis mencoba memberikan solusi dengan cara pencarian tidak hanya didasarkan oleh isi dokumen halaman web tetapi juga adanya rekomendasi dari penulis dokumen halaman web lain. Semakin banyak rekomendasi dari penulis lain maka dokumen halaman web tersebut dikatakan semakin relevan terhadap query inputan yang diberikan oleh user.

Terdapat beberapa jenis algoritma yang menggunakan prinsip link – based analysis. Salah satu algoritma tersebut ialah algoritma SALSA. Algoritma SALSA menekankan pembentukan graph menjadi bipartite graph yaitu satu untuk graph hub dan satu untuk graph authority. Hub ialah nilai yang dimiliki suatu dokumen karena menuju dokumen lain sementara authority ialah nilai yang dimiliki oleh suatu dokumen karena dituju oleh dokumen lain. Dari hasil pengujian Tugas Akhir ini, Algoritma SALSA memberikan hasil yang cukup baik berdasarkan nilai precision dan IAP-nya. Hal ini dapat dilihat bahwa rata-rata dokumen halaman web yang memiliki jumlah backlink yang cukup banyak akan memiliki peringkat pengurutan yang jauh lebih baik daripada dokumen halaman web yang memiliki jumlah backlink yang sedikit.

Kata Kunci : authority, backlink, dokumen, hub, link-based analysis, text-based analysis, SALSA algorithm.

Abstract

The process of searching and sorting documents on the web pages in search engine is divided into two parts, namely by using a text-based analysis and link-based analysis. The text-based analysis process is the ranking process of web page documents based on existing text in the documents. Search process by using this concept has some shortcomings such as to search for relevant documents in accordance with the user desires, the user is expected to enter a query as precisely as possible with the existing terms in the document.

Search by using the concept of link-based analysis tries to provide a solution by way of searches based not only on the contents of web page documents but also the recommendation from the authors of other web page documents. The more recommendations from other authors, the more relevant the web page document to the query given by user.

There are several types of algorithms that use the principles of links - based analysis. One of these algorithms is the SALSA algorithm. SALSA algorithm emphasizes the formation of a graph into bipartite graph, i.e. one for hub graph and the other for the authority graph. Hub is the value owned by a document for it addresses other document while the authority is the value owned by a document for it is addressed by other document. From the results of this final testing, SALSA algorithm gives good results based on the value of its precision and IAP. This can be seen that, on the average, web page documents that have a rather large number of backlinks will have the higher rank than those that have less number of backlinks.

Keywords : authority, backlink, document, hub, link-based analysis, text-based analysis, SALSA algorithm.

1. Pendahuluan

1.1 Latar belakang

Pencarian dengan menggunakan mesin pencari telah menjadi hal yang umum dan sangat dibutuhkan oleh masyarakat pada saat ini. Pencarian dilakukan dan diharapkan menghasilkan hasil yang jauh lebih tepat dan akurat. Pencarian tidak lagi hanya didasarkan oleh kemiripan *query* tetapi juga mulai didasari atas adanya rekomendasi dari penulis dokumen halaman *web* yang lain.

Proses pencarian berdasarkan prinsip *text-based analysis* didasari atas isi teks yang ada di dalam suatu dokumen [8]. Proses pencarian berdasarkan prinsip ini memiliki kekurangan yaitu untuk mencari dokumen yang relevan maka *user* dituntut untuk memberikan inputan *query* yang secocok mungkin dengan *term* yang ada di dalam dokumen. Oleh karena itu, munculnya mesin pencari dengan menggunakan prinsip *link analysis*. Prinsip pencarian berdasarkan *link analysis* ialah proses pencarian dan pengurutan dokumen halaman *web* berdasarkan informasi yang ada di dalam *link* tersebut atau disebut juga adanya rekomendasi dari penulis dokumen halaman *web* yang lain [8].

Algoritma SALSA adalah algoritma yang akan dibahas dalam Tugas Akhir ini. Algoritma SALSA(*Stochastic Approach for Link-Structure Analysis*) merupakan salah satu jenis algoritma *link analysis*. Algoritma ini dikembangkan oleh Lempen dan Moran dimana dalam menentukan halaman web yang *relevan* menggunakan prinsip *hub* atau *authority* [1,3,8,10]. *Authority* ialah nilai yang dimiliki oleh suatu dokumen karena dituju oleh dokumen halaman *web* lain [1,3,8,10]. *Hub* ialah nilai yang dimiliki suatu dokumen karena mengacu ke dokumen lain [1,3,8,10]. Penentuan pengurutan tertinggi didasari atas nilai *authority* dimana semakin besar nilai *authority*-nya maka semakin tinggi pengurutannya.

1.2 Perumusan masalah

Adapun masalah yang akan diselesaikan pada Tugas Akhir ini, yaitu:

1. Bagaimana pengaruh jumlah *link* yang dimiliki oleh suatu dokumen terhadap pengurutan ke*relevan*annya?
2. Bagaimana pengaruh jumlah dokumen yang ter-*crawler* untuk menentukan pengurutan dokumen – dokumen halaman *web*?
3. Bagaimana pengaruh *single document* dan *multi document* dalam proses pengurutan dengan menggunakan Algoritma SALSA?
4. Bagaimana pengaruh jumlah *root set* dalam pengurutan dokumen halaman *web*?

1.3 Batasan Masalah

Adapun batasan-batasan yang diberikan dalam penyelesaian masalah Tugas Akhir ini adalah sebagai berikut :

Batasan-batasan masalah pada Tugas Akhir ini adalah :

1. Penelitian pada Tugas Akhir ini berfokus pada Algoritma SALSA(*Stochastic Approach for Link-Structure Analysis*).
2. Proses *crawling* dilakukan secara *On-line*.

3. Proses perhitungan Algoritma SALSA terhadap dokumen ter-*crawling* dilakukan secara *offline*.
4. Sistem tidak menangani pencarian *root set* berdasarkan *text-based analysis*.
5. Sistem tidak menangani pembobotan *term* pada proses *searching*.

1.4 Tujuan

Tujuan Tugas Akhir ini adalah sebagai berikut.

1. Menganalisis pengaruh jumlah *link* yang dimiliki oleh suatu dokumen halaman *web* dalam pengurutan dokumen halaman *web* tersebut.
2. Menganalisis pengaruh banyaknya jumlah dokumen yang ter*crawling* dalam penentuan pengurutan dokumen halaman *web*.
3. Menganalisis pengaruh *single* dan *multi document* terhadap pengurutan dokumen halaman *web*.
4. Menganalisis pengaruh banyaknya jumlah *root set* terhadap hasil pengurutan dokumen halaman *web*.

1.5 Metodologi penyelesaian masalah

Metode yang digunakan untuk menyelesaikan permasalahan-permasalahan Tugas Akhir ini terdiri dari langkah-langkah sebagai berikut.

1. Pengumpulan data dan studi literatur
   Mempelajari dan memahami salah satu algoritma *link analysis yaitu* algoritma SALSA melalui literatur berupa buku, makalah, atau jurnal dari berbagai media terutama Internet.
2. Pembangunan Aplikasi
   Pembangunan aplikasi yang meliputi:
   a. Perancangan Sistem
      Menyiapkan dokumen yang akan digunakan utnuk dilakukan proses pengurutan berdasarkan algoritma SALSA. Proses selanjutnya dilakukan perancangan sistem yang nantinya dapat menghasilkan suatu sistem yang dapat melakukan pengurutan dokumen halaman *web* berdasarkan jumlah *link* yang dimiliki oleh dokumen halaman *web* tersebut.
   b. Implementasi
      Di tahap ini dilakukan implementasi algoritma SALSA pada dokumen yang diperoleh, yang meliputi:
      - Memproses sejumlah *root set* yang diberikan untuk dilakukan proses *crawling* terhadap sejumlah *root set* tersebut.
      - Setiap *root set* yang di*crawling* dilakukan proses penyimpanan *link* yang dimiliki oleh *root set* tersebut.
      - Proses perhitungan jumlah *link* yang ter*crawling* berdasarkan algoritma SALSA.
   c. Pengujian
      Sistem aplikasi yang sudah dibangun kemudian diuji untuk mengetahui apakah sistem sudah berjalan seperti yang diharapkan.

1) Untuk menganalisis efektifitas aplikasi, pengujian dilakukan dengan menjalankan aplikasi *search engine* dan kemudian dilakukan percobaan dengan menggunakan *single document* dan *multi document*.
2) Untuk menganalisis pengaruh jumlah dokumen dan jumlah *root set* maka dilakukan pengujian dengan memperbanyak jumlah atau nilai keduanya apakah dapat menghasilkan hasil pengurutan yang berbeda.

d. Analisis Hasil
Pada tahap ini didapat *output* dari hasil perhitungan algoritma SALSA kemudian dihitung nilai *IAP* dan *precision* berdasarkan prinsip *backlink metric* dan dianalisis apakah algoritma SALSA tepat untuk proses perangkingan.

e. Penyusunan Laporan
Pada tahap ini, akan dilakukan penyusunan laporan akhir sekaligus dokumentasi dengan mengikuti kaidah penulisan yang benar dan sesuai dengan ketentuan yang ditetapkan oleh institusi.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari hasil pengujian dan analisis yang telah dilakukan sebelumnya, yaitu :

1. Algoritma *SALSA* kurang sesuai dilakukan terhadap data dengan tipe *single document* dan lebih cocok untuk tipe *multi document* karena proses iterasi pada *single document* tidak dapat dilakukan.
2. Semakin besar jumlah *root set* dan jumlah dokumen maka semakin baik perhitungan nilai *precision* dan IAP karena jumlah dokumen halaman relevan yang ter-*retrieved* semakin banyak.

5.2 Saran

1. Adanya pembobotan *query* dan *term document* agar hasil yang didapat lebih maksimal.
2. Sistem menggabungkan perankingan secara *text-based* dan *link-based* dalam perankingannya.

Daftar Pustaka

[1]	Borodin Allan, Roberts, O Gareth , Rosenthal Jeffrey S, Tsaparas Panayiotis. Link Analysis Ranking : Algorithms, Theory, and Experiments. www.cs.toronto.edu/~tsap/publications/hubs-journal.ps Diakses pada 11 Maret 2010 pukul 11.00 WIB.
[2]	Budianto, Arifin Zainal Agus, Lili Suhadi. Perancangan dan Pembuatan Perangkat Lunak Penelusur Web (Web Crawler) Menggunakan Algoritma Pagerank. Teknik Informatika, Institut Teknologi Sepuluh November Surabaya, 2003. http://www.its.ac.id/personal/files/material/1261-agusza-webcrawler.pdf. Diakses pada 10 Agustus 2010 pukul 15.00 WIB.
[3]	Farahat, Ayman, Lofaro Thomas, Miller Joel C, Rae Gregory, Ward Lesley A. Authority Rankings From HITS, PAGERANK, AND SALSA : Existence, Uniqueness, and Effect of Initialization. http://www.math.hmc.edu/~ward/paperpdfs/hitsheaderbw6Jan05.pdf. Diakses pada 18 Oktober 2009 pukul 20.13 WIB.
[4]	Firdaus, Yanuar . "Introduction to Information Retrieval", Slide Kuliah, Institut Teknologi Telkom Bandung,Mei 2008
[5]	Firdaus, Yanuar . "Web Search", Slide Kuliah, Institut Teknologi Telkom Bandung,Mei 2008
[6]	Henzinger, Monika . "Link Analysis in Web Information Retrieval" http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/9019.pdf
[7]	Irvine, Link Analysis www.**powerpoint**-search.com/**link-ppt**.html Diakses pada 6 Oktober 2009 pukul 13.00 WIB
[8]	Lempel R, Moran S. SALSA: The Stochastic Approach for Link Structure Analysis. http://www.cparity.com/projects/AcmClassification/samples/383041.pdf. Diakses pada 18 Agustus 2010 pukul 20.19 WIB.
[9]	Precision and Recall. http://en.wikipedia.org/wiki/Precision_%28information_retrieval%29. Diakses pada 15 Juli 2010 pukul 15.30 WIB.
[10]	Signorini Alessio. A survey Ranking Algorithms. http://www.cs.uiowa.edu/~cremer/courses/cs2/SignoriniRankingAlgSurvey.pdf. Diakses pada 19 Agustus 2010 pukul 20.22 WIB.